An Epistemic Justification of the Law of Total Probability

by

David A. Lane
University of Minnesota
Technical Report No. 491
July, 1987

Text of a lecture presented at the First International Congress of the Bernoulli Society, Tashkent, U.S.S.R., September 1986.

AN EPISTEMIC JUSTIFICATION OF THE LAW OF TOTAL PROBABILITY

David A. Lane
School of Statistics, University of Minnesota, Minneapolis, MN U.S.A.

INTRODUCTION: KOLMOGOROV'S AXIOMS AND EPISTEMIC PROBABILITY

The publication of Kolmogorov's Foundations of the Theory of Probability in 1933 was a decisive event in the development of probability theory as a mathematical discipline. Kolmogorov's six axioms identified probability as a special case of the measure and integration theory developed by Lebesgue, and so the powerful machinery of that theory could be used to establish probabilistic results rigorously and in great generality. However, Kolmogorov left open the question of how to interpret the mathematical entities Ω, $\mathcal{F}$ and P to which his axioms referred (except for finite Ω, where he seemed to regard the elements of Ω as descriptions of the possible outcomes of some repeatable experiment, and P(A) the proportion of times that the outcomes in A had occurred in some finite sequence of repetitions of the experiment). In addition, Kolmogorov acknowledged that some elements of his axiomatization--in particular, the continuity axiom, Axiom VI--introduced "arbitrary limitations", not required by the nature of whatever it is that the mathematical entities actually represent.

As Ian Hacking has argued in his book The Emergence of Probability, two different notions of the nature of probability have coexisted, sometimes uneasily, since mathematicians first turned their attention to probabilistic problems. Epistemic probability describes degree of belief, while aleatory probability describes an attribute of objects: the propensity of what Kolmogorov (1956) refers to as "observable random phenomena" to achieve stable relative frequencies under extended series of independent repetitions.

Most workers who have adopted the Kolmogorov axiomatization, including Kolmogorov himself, have emphasized the aleatory interpre-

tation of probability. However, there seems to be an increasing recognition by statisticians that epistemic interpretations are also useful, if not necessary, for two purposes that are fundamental to inference: stating predictions about future observables, in such a way that the predictions are not conditional on unobservable "states of the world"; and assigning probabilities to propositions about such unobservable "states of the world". Indeed, Bayesian statistical inference consists precisely in the utilization of epistemic probability for these two purposes.

Are Kolmogorov's axioms warranted if probability is interpreted epistemically--that is, as a measure of degree of belief? Many authors have considered this question, taking many different approaches to the problem of defining "degree of belief." Virtually all of the elements of Kolmogorov's axiomatization have been challenged somewhere in this literature. Some authors--in particular Koopman (1940), Good (1950) and Smith (1961)--have argued for qualitative or set-valued measures for belief. Others—for example, de Finetti (1974) and de Jouvenal (1967)--have accepted a real-valued measure, but have denied the appropriateness of Ω (that is, an exhaustive list of nondecomposable propositions); while others have granted Ω, but have challenged the requirement that $\mathcal{F}$ should be a σ-field or even a field (see, for example, Fine (1973)). As far as P is concerned, the epistemic justification for its additivity has been questioned, for example by Shafer (1976), whose theory of evidence has been much admired recently by some workers in artificial intelligence expert systems. Others, especially de Finetti (1974) and Savage (1954), have justified finite additivity, but rejected the necessity of countable additivity. Finally, many authors, including Jeffreys (1961) and Renyi (1956) and, more recently, Reggazini (1983), Holzer (1985) and Armstrong and Sudderth (1985), have found Kolmogorov's normalization requirement that $P(\Omega)$ should equal 1 to be too restrictive to permit all reasonable comparisons of uncertainty.

I will not attempt to summarize all this literature in this talk. Rather, I will focus on one approach to epistemic probability and its bearing on a question of fundamental importance, the logical validity of the Law of Total Probability. A major theme will be the

interplay between the interpretation of probability and its appropriate mathematization.

COHERENCE AND COUNTABLE ADDITIVITY

Around the same time that Kolmogorov formulated his axioms for probability theory, de Finetti was developing a quite different approach to the subject. De Finetti was concerned with the problem of how to act in the face of uncertainty. He argued that it was necessary for an individual to measure his uncertainty about the truth of a proposition in terms of his propensity to act as though the proposition were true, and so he proposed a definition of probability in terms of a decision with simple economic consequences. Specifically, the probability of a proposition is the price at which the assessor is neutral between buying and selling a ticket that is worth $1 if the proposition is true and is otherwise worthless.

With this definition, de Finetti was able to give a precise meaning to consistent reasoning about uncertainty. Suppose an individual measures his uncertainty about a set of propositions, so that he has, according to de Finetti's definition, determined the price of a corresponding set of tickets. Could someone transact with the assessor to buy or sell some of these tickets, at the assessor's prices, in such a way that the assessor must pay out more than he receives, no matter which of the propositions turn out to be true and which to be false? If so, the set of assessments is incoherent: the possibility of sure loss concretizes the inconsistent reasoning underlying the assessments. A set of assessments that cannot result in sure loss is called coherent.

To make the definition of coherence operational, de Finetti insisted that the transactions described there be limited to a finite number of sales or purchases. It is possible to make an infinite number of probability assessments with the stroke of a pen, but there is no physically realizable way to exchange more than a finite number of dollars.

De Finetti proved that a set of probability assessments on some collection of subsets of a set Ω is coherent if and only if it is

consistent with a finitely additive probability measure defined on all subsets of Ω. Thus, de Finetti rejected Kolmogorov's Axiom VI, which requires probability measures to be countably additive. In his writings, he has given a number of examples of epistemically justifiable noncountably additive probability distributions, including the uniform distribution on the integers (see also Hill (1980) and Scozzafava (1981), which presents an example in the context of the first digit problem).

THE LAW OF TOTAL PROBABILITY

The Law of Total Probability plays a fundamental role in probability assessment. To determine the probability of a proposition A, an assessor often finds it convenient to "extend the conversation" (in Dennis Lindley's phrase) to include a set Π of mutually exclusive, exhaustive propositions; then to evaluate a probability distribution μ over Π and, for each h in Π, the conditional probability $P(A|h)$; and finally to assess the probability for A by the formula

$$(1) \qquad P(A) = \int P(A|h)\,d\mu(h).$$

In Foundations of Probability Theory, Kolmogorov defined $P(A|h)$, when h has positive probability, as the ratio of the two unconditional probabilities $P(Ah)$ and $P(h)$, so that (1) is true by definition (and the additivity of P) in this case. However, since he offered no interpretation of conditional probability that could be used to produce a direct assessment for $P(A|h)$, there is no way to determine $P(A|h)$ except in terms of unconditional probability assessments, and so the assessment strategy described above is meaningless. In case Π is uncountable, Kolgomorov defined the random variable $P(A|\cdot)$ in terms of a Radon-Nikodym derivative, again in such a way as to make (1) true by definition. But $P(A|\cdot)$ is not even defined pointwise on the set of elements of Π that have probability zero, so the assessment strategy is impossible to implement in this case as well.

In contrast to the Kolmogorov approach, de Finetti directly defined conditional probability. For two propositions A and B, the conditional probability $P(A|B)$ is the price at which the assessor is neutral between buying and selling a ticket that is worth \$1 if A

and B are both true, worth nothing if B is true and A is false and for which the purchase price is refunded if it turns out that B is false. With this definition, de Finetti proved that unconditional assessments P(AB) and P(B) and conditional assessment P(A|B) were coherent if and only if P(AB) equalled the product of P(B) and P(A|B) (so that the relation between conditional and unconditional probability that Kolmogorov introduced by definition for P(B) > 0 can be obtained as a consequence of de Finetti's coherence requirement).

If Π is a finite partition, and the probabilities P(A) and $\{P(A|h)\}_{h \text{ in } \Pi}$ are assessed, coherence requires that equation (1) must hold. However, if Π is infinite, de Finetti argued that (1) need not hold, because of the limitation to a finite number of cash transactions in the definition of coherence. Thus, an assessor could coherently violate the Law of Total Probability.

Now consider the validity of the Law of Total Probability, not for a particular proposition A, but for the entire distribution P. That is, suppose $\mathcal{F}$ and $\mathcal{G}$ are algebras of subsets of Ω, with $\mathcal{F}$ contained in $\mathcal{G}$, and Π a $\mathcal{G}$-measurable partition of Ω. Suppose further that P is a probability measure on $(\Omega,\mathcal{G})$ with Π-marginal μ, and for each h in Π, $P(\cdot|h)$ is a probability measure on $(\Omega,\mathcal{F})$ with support in h. Then

(a) P is Π-disintegrable with disintegration $\{P(\cdot|h)\}$ if equation (1) holds for all A in $\mathcal{F}$.

(b) P is conglomerable with respect to $\{P(\cdot|h)\}$ if for all A in $\mathcal{F}$

$$\inf_{h \text{ in } \Pi} P(A|h) \leq P(A) \leq \sup_{h \text{ in } \Pi} P(A|h). \tag{2}$$

If (2) fails for some A in $\mathcal{F}$, then P is nonconglomerable. (Clearly, if P is nonconglomerable, it cannot be disintegrable.)

Even the Kolmogorov theory admits situations in which the Law of Total Probability fails to hold for all events in an algebra $\mathcal{F}$: that is, as is well-known, there are examples of countably additive distributions P that are not disintegrable with respect to uncountable partitions (see Blackwell and Dubins (1975)). The situation is worse with respect to noncountably additive distributions. It is easy to find examples, like the following, of noncountably additive distributions that are nonconglomerable with respect to countable partitions,

which cannot happen if P is countably additive.

Example: Suppose P_1 and P_2 are probability measures on the nonnegative integers: P_1 is countably additive, supported by the even integers, and gives each even integer positive probability; P_2 is a diffuse distribution (that is, $P_2\{i\} = 0$ for each integer i) supported by the odd integers. Let $P = (P_1+P_2)/2$, and define the partition with elements $h_i = \{2i,2i+1\}$ for $i = 0,1...$ To satisfy the Law of Total Probability for the singleton $\{2i\}$, $P(\{2i\}|h_i)$ must equal 1 for each i. With this assessment, the Law then fails for any set containing a subset of the odd integers with positive P_2-probability.

In fact, nonconglomerability with respect to countable partitions characterizes noncountable additivity. De Finetti (1974) conjectured and Schervish, Seidenfeld and Kadane (1984) and Hill and Lane (1985) proved that for any noncountably additive probability measure P, there must be a countable partition Π, such that P is nonconglomerable with respect to $\{P(\cdot|h)\}_{h \text{ in } \Pi}$, when conditional probability measures are defined with respect to all nonempty measurable sets. This connection between noncountable additivity and nonconglomerability has inclined several statisticians who subscribe to de Finetti's foundational framework, and so accord an inferential role to noncountably additive distributions, to accept nonconglomerable assessments as a basis for inference as well (for discussion and examples, see de Finetti (1972), Hill (1980), Scozzafava (1984) and Kadane, Schervish and Seidenfeld (1986)). The position is challenged in the next two sections.

PROSPECTIVE PROBABILITIES

De Finetti's result that distributions that are nonconglomerable with respect to infinite partitions are coherent is mathematically unarguable. The question is whether the way in which he defined conditional probability corresponds to any inferential problem of interest. That is, do the rules for the economic transactions incorporated in that definition adequately translate the epistemic role that conditional probability plays in inference?

It is a truism to say that all epistemic probabilities are conditional, on everything that the assessor believes to be true (or that

he chooses to act as though he believes to be true). The conditional probabilities that figure in de Finetti's definition have a different character. They are evaluated as though the evaluator believes that the conditioning proposition is true, even though he explicitly asserts, through his unconditional probability assessments, that this is not the case.

What is the point of evaluating such quantities? Since subjective probability assessment is never easy, it makes no sense to evaluate probabilities that do not figure in particular inferential problems, and even if they are evaluated, it seems pointless to test whether they are consistent with evaluations that do have an inferential role to play. I know only two kinds of situations in which it is inferentially advantageous to evaluate a probability conditionally on a proposition whose truth the evaluator is unwilling to assume. The first, discussed in this section, has to do with updating opinion through time, and the second, discussed in the next section, involves the use of parametric models to assess probabilities for propositions that are not directly accessible to the evaluator's knowledge and experience.

Suppose the assessor believes that tomorrow he will observe a quantity X whose value is informative about another quantity, Y. In addition, he believes that nothing else he will learn in the interim should affect his opinions about Y, currently encoded in a probability distribution P. In this situation, he may choose to assess distributions $P(\cdot|X=x)$ for each possible value of the quantity X. The purpose of these assessments is to discipline and direct his response to the actual value of X when he learns it tomorrow: $P(\cdot|X=x)$ represents an opinion about Y that he currently contracts to adopt tomorrow, should he observe x, and should his belief that he will learn nothing else relevant about Y turn out to be true.

Why should he agree now to specify his future opinions? Because he knows how easy it is for the shock of the new to overwhelm previously accrued information bearing on Y, as encoded in P, and he believes that the distancing from that shock that he achieves through the prospective assessment, considering each candidate new observation in turn, can help him achieve a successful reconciliation between

old and new information. (Evaluations of this sort can also help the assessor to decide whether it is worth observing X at all, in terms of its effects on his opinions about Y; but in this "design" context also, $P(\cdot|X=x)$ must be interpreted as a contract on a future opinion. In addition, these distributions, along with P, can also be assessed retrospectively, if the assessor wants to guard against overreacting to an observation he has just made. Carrying out such a retrospective analysis is a reasonable response when something other than X is learned in the interim, with X and that "something else" playing the role that X plays in the discussion here.)

If we regard $P(A|X=x)$ as a contract to adopt a particular opinion about the proposition A (referring to the value of Y) if and when X turns out to equal x, what should be the economic interpretation for this quantity? Today, before X is observed, the ticket priced at $P(A|X=x)$ can be reserved with no cash payment (since the opinion it represents is not in force today); if it is reserved, and $X=x$ tomorrow, the opinion is in effect and so the price must be turned over to the seller, in which case if A turns out to be true, the seller must pay the buyer \$1 and otherwise nothing; while if the ticket is reserved and $X\neq x$ tomorrow, the reservation is cancelled and no money changes hands. That is: tickets are actually purchased only when the opinions they represent are in effect; an assessor willing to commit himself to an opinion in advance will accept reservations for the future purchase or sale of tickets at his pre-specified price. I claim that this formulation captures in economic terms exactly what the assessor means when he evaluates $P(A|X=x)$ prior to observing X.

With this formulation, the Law of Total Probability must obtain. That is, suppose the assessor evaluates P on $\mathcal{F}$ (a set of propositions about Y), and, for each possible value x for X, $P(\cdot|X=x)$ on $\mathcal{F}$. In the economic test for coherence, the bettor is allowed to buy a finite number of tickets from the P-book; and, for each x, to contract for a finite number of tickets from the $P(\cdot|X=x)$-book--these contracts become operative when and if x is observed. Because only one x can be observed, such a system of contracts, purchases and sales involves only a finite number of cash transactions. Unless the assessments

satisfy equation (1) for some finitely additive distribution μ on the values of X and all propositions A in $\mathcal{F}$, there exists a system of contracts, purchases and sales as described above that will guarantee the assessor a loss of at least \$c (for any prespecified amount c), no matter which x is observed or which of the propositions in $\mathcal{F}$ are true. Thus, in this sequential learning context, assessments that violate the Law of Total Probability are incoherent. This assertion is proved in Lane and Sudderth (1984) and Lane and Sudderth (1985).

PUTATIVE PROBABILITIES AND STATISTICAL MODELS

Suppose $\mathcal{F}$ is a set of observable propositions whose probabilities must be evaluated for some inferential problem, but are difficult to evaluate directly. Suppose also that Θ is a set of propositions that the assessor believes to be mutally inconsistent and exhaustive, and for each θ in Θ and A in $\mathcal{F}$, $P(A|\theta)$ is directly accessible to the assessor's knowledge and experience and so easy to assess. Suppose also that which θ is true is unobservable, so the interpretation of conditional probability as a contract discussed above could not be made operational (when would the money change hands?). Rather, $P(A|\theta)$ has a merely putative meaning: if the evaluator lived in a world in which θ is true and is otherwise like his own, what would be his degree of belief in A? That is, while he is confused about the propositions in $\mathcal{F}$ in the world he actually inhabits, the assessor can imagine how he would believe from the vantage points offered by the elements of Θ.

Now suppose that the assessor also evaluates his probability distribution P on the elements of $\mathcal{F}$ directly (difficult though this task may be). What relation need these evaluations bear to those conditional on the elements of Θ? Suppose a bettor is allowed to purchase or sell a finite number of P-priced tickets. Since the assessor is confident about his ability to evaluate probabilities for the elements of $\mathcal{F}$ from every θ-vantage point, and he believes that one of these vantage points gives a view of the world he actually inhabits, he may evaluate the worth of the bettor's transaction with him from each θ-vantage point in turn. That is, if L represents his

loss from the transaction, he can evaluate with confidence $E(L|\theta)$ for each θ in Θ. If each of these is greater than \$c, then, whichever θ, coupled with the rest of his general background information, actually describes the world in which the assessor lives, from the vantage point of that world the assessor faces a loss of at least \$c. This anticipated loss represents an inconsistency between his direct assessment P and his assessments from the vantage points in Θ. Lane and Sudderth (1984) prove that this anticipated loss can be avoided if and only if there is a finitely additive probability distribution μ on Θ such that $P(A) = \int P(A|\theta)d\mu(\theta)$ for all A in $\mathcal{F}$; that is, if the (difficult) unconditional assessment P is linked to the (accessible) conditional assessments $\{P(\cdot|\theta)\}$ by the Law of Total Probability.

This result responds to a question raised by Piccinato (1980). He asked "How to deal with statistical models if we must get rid of conglomerability? Of course the ground for accepting or refusing conglomerability (or complete additivity) is a logical one, and must be independent from the mentioned question... For 'statistical model' I mean as usual a set of probability distributions on the space of possible outcomes, possibly indexed by a parameter whose actual value is unknown... Can we live with nonconglomerability without giving up statistical models?" Of course, some statistical models have attributes like the set Θ discussed above: the elements in the parameter space can be interpreted as propositions about the world, one of which the observer believes to be true, and it is possible to confidently assess putative probabilities for observables from the vantage point of each element of the model. When these conditions are satisfied, the logical ground that Piccanato seeks is the relevance of the evaluations of the anticipated loss from the vantage point of each parameter value as defined above; to avoid that loss, conglomerability must hold for the probabilities of observables with respect to their putative probabilities evaluated given parameter values.

SOME OBJECTIONS AND ANSWERS

In the previous two sections, I have argued that in the situations in which conditional probabilities play a role in inference, as prospective probabilities and putative probabilities, their evaluation

should respect the Law of Total Probability with respect to finitely additive distributions μ. This point of view is criticized by Kadane, Schervish and Seidenfeld (1986), who take the position that "...the attempt to modulate nonconglomerability is misguided. If nonglomerability is necessary for 'consistency', then nothing less than countably additive distributions suffice with denumerable partitions, and even 'proper' priors may suffer nonconglomerability in non-denumerable partitions..." I disagree with their position for two reasons. First, I can think of no inferential problem in which it is reasonable to evaluate probabilities conditional on all possible events, and I believe that consistency tests should be applied only to the probabilities that play a role in particular inferential problems. Secondly, I do not understand the relevance of the "called-off bet" economic formulation of conditional probability, with purchase price handed over before the opinion represented by the conditional probability is relevant. Rather, I believe that the actual role that conditional probabilities play in inference should be the basis of their economic interpretation, and I have attempted to isolate two such roles and provide their respective (and different) interpretations in the two previous sections.

Several authors have objected to the criterion of anticipated loss as a condition for coherence developed in the previous section. I share this objection when it is applied to situations in which a statistical model has been adopted "for convenience", with no presumption that probabilities given θ really describe the evaluator's putative beliefs in a world he believes might be his own and in which he can comfortably measure his uncertainty. In such cases, I cannot interpret probabilities given θ at all, much less decide how consistent they are with probabilities that actually do measure some of the evaluator's beliefs. But if putative probabilities are available as described in the previous section, then evaluating expectations from the same vantage point seems reasonable and meaningful; and if I face a \$c loss in all possible worlds, then in particular I do <u>in this one</u> as well. (It is important to realize that these expectations need bear no interpretation in terms of "repeated sampling", as Scozzafava (1984) seems to imply; they are just assessable previsions in an

imagined world.)

REFERENCES

Armstrong T., Sudderth W. (1985). Locally coherent rates of exchange. Technical Report University of Minnesota School of Statistics No. 459.

Blackwell D., Dubins L. (1975). On existence and nonexistence of proper, regular, conditional distributions. Annals of Probability 3, 741-752.

de Finetti B. (1972). Probability, Induction and Statistics. New York, Wiley.

de Finetti B. (1974). Theory of Probability. New York, Wiley. (Translation of 1970 Italian original).

de Jouvenal B. (1967). The Art of Conjecture. New York: Basic Books.

Fine T. (1973). Theories of Probability - An Examination of Foundations. New York, Academic Press.

Good I.J. (1950). Probability and the Weighing of Evidence. London, Griffin.

Hacking I. (1975). The Emergence of Probability. London, Cambridge.

Hill B. (1980). On some statistical paradoxes and nonconglomerability. In: Proceedings of the First International Meeting in Bayesian Statistics. Valencia, University Press, 39-66.

Hill B., Lane D. (1985). Conglomerability and countable additivity. Sankhya 47, 366-379.

Holzer S. (1985). On coherence and conditional prevision. Analisi Funzionale e Applicationi, Serie VI, Vol. IV-C, no. 1.

Jeffreys H. (1961). Theory of Probability (Third Ed.). Oxford, Oxford Univ. Press.

Kadane J., Schervish M., Seidenfeld T. (1986). Statistical implications of finitely additive probability. In: Bayesian Inference and Decision Techniques, Zellner, A., Goel, P. (Ed.) Amsterdam, Elsevier, 59-76.

Kolmogorov A. (1956). Foundations of the Theory of Probability. New York, Chelsea Publishing Co. (Translation of 1933 German original).

Koopman B. (1940). The bases of probability. Bull. Amer. Math. Soc, 46, 763-774.

Lane D., Sudderth W. (1984). Coherent predictive inference. Sankhya 46, 166-185.

Lane D., Sudderth W. (1985). Coherent predictions are strategic. Annals of Statistics 13, 1244-1248.

Piccanato L. (1980). Comment on Hill (1980). In: Proceedings of the First International Meeting in Bayesian Statistics. Valencia, University Press, 51-53.

Reggazini E. (1983). Coherent conditional probabilities, finite additivity, extensions. Tech. Report Dipart Sc. Statist., Univ. of Bologna.

Renyi A. (1956). On conditional probability spaces generated by a dimensionally ordered set of measures. Theory of Prob. and its Applications 1, 61-71.

Savage L. (1954). Foundations of Statistics. New York, Wiley.

Schervish M., Seidenfeld T., Kadane J. (1984). The extent of nonconglomerability of finitely additive probabilities. Z.

Wahrscheinlichkeitstheor. verw. Geb. 66, 204-226.
Scozzafava R. (1981). Nonconglomerability and the first digit problem. Statistica 41, 561-565.
Scozzafava R. (1984). A survey of some common misunderstandings concerning the role and meaning of finitely additive probabilities in statistical inference. Statistica 44, 21-45.
Shafer G. (1976). A Mathematical Theory of Evidence. Princeton, Princeton University Press.
Smith C. (1961). Consistency in statistical inference and decision. J.R.S.S.(B) 23, 1-37.